LANGUAGE SYSTEMS, INC.
MUC-4 TEST RESULTS AND ANALYSIS[1]

Christine A. Montgomery
Bonnie Glover Stalls
Robert E. Stumberger
Naicong Li
Robert S. Belvin
Alfredo Arnaiz
Susan B. Hirsh

Language Systems, Inc.

6269 Variel Avenue, Suite F
Woodland Hills, CA 91367
(818) 703-5034

Internet: chris@lsi.com

NLP OBJECTIVES

LSI's overall natural language processing (NLP) objective is the development of a broad coverage, reusable system which is readily transportable to additional domains, applications, and sublanguages in English, as well as providing a foundation for our multilingual work. Our system, called DBG, for Data Base Generator, is comprised of a set of NLP components which have been developed, extended, and rebuilt over a period of some years. The core of the system is an innovative Principle-based parser, using ideas from [1], which we began developing in the course of MUC-3 to replace our previous chart parser. Our approach thus relies on the concept of powerful, robust parsing as the most crucial component in an NLP system. In applying our NLP system to text extraction, our ultimate objective is to develop a high quality text extraction system, where "high quality" is defined as scoring above 80% -- a number well beyond any current MUC scores.

In line with these NLP objectives, our major focus for MUC-4 was a follow-up to our main "lesson learned" in MUC-3, which was to acquire a machine-readable dictionary (MRD) and integrate its content into the DBG system. When attempts to acquire the computer-friendly Longmans or one of the Oxford Dictionaries were unsuccessful, we turned to ACL's CD-ROM containing the Collins English Dictionary. The most correct version of the CED on the ACL CD-ROM was apparently developed directly from a medium prepared for the typographer, and unfortunately lacks any documentation of features, fonts, language, etc. The effort of acquiring and integrating the CED was clearly a worthwhile endeavor, since we were able to increase the number of entries in our lexicon three-fold in a relatively short time (see Table 1). The increase in lexicon size will benefit all the applications LSI is currently working on.

1. The work reported in this paper was supported in part by the Defense Advanced Research Projects Agency, Software and Intelligent Systems Technology Office, under Contract No. N66001-90-C-0192 (Subcontract 19-930042-31 to SAIC), and by the U. S. Army Ballistic Research Laboratory under Contract No. DAAA15-89-C-0004 (Subcontract No. 05-562-01 to Logicon, Inc.)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1992		**00-00-1992 to 00-00-1992**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
MUC-4 Test Results and Analysis	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language Systems Inc,6269 Variel Avenue,Woodland Hills,CA,91367**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**5**	

	MUC3	MUC4
STEMS		15,285
INFLECTED FORMS		14,561
TOTALS	est at 10,000	29,846

Table 1. LSI Lexicon Statistics

RESULTS

The complete LSI TST3 and TST4 score reports are included in Appendix G, "Final Test Score Summaries". As an indication of system development during MUC4, we can compare our TST3 results with our results on the MUC-4 interim test (TST2). The relevant portions of the TST3 and TST2 results are shown in Tables 2 and 3. Figure 1 graphically presents the TST2 and TST3 recall and precision matrices.

SLOT	POS	ACT	COR	PAR	INC	ICR	IPA	SPU	MIS	NON	REC	PRE	OVG
inc-total	385	309	16	7	18	1	1	283	359	255	5	6	92
perp-total	230	87	3	2	3	0	2	79	222	351	2	4	91
phys-tgt-total	195	68	0	0	3	0	0	65	192	728	0	0	96
hum-tgt-total	449	63	20	4	2	0	4	37	423	772	5	35	59
Matched/Missing	1259	71	39	13	11	1	7	8	1196	898	4	64	11
Matched/Spurious	113	527	39	13	26	1	7	464	50	1277	40	9	88
Matched Only	113	71	39	13	11	1	7	8	50	69	40	64	11
All Templates	1259	527	39	13	26	1	7	464	1196	2106	4	9	88
Set Fills Only	593	36	26	4	4	0	2	2	559	418	5	78	6
String Fills Only	337	9	5	2	2	0	2	0	328	257	2	67	0

Table 2. TST2 (MUC4 Interim Test) Summary Scores

SLOT	POS	ACT	COR	PAR	INC	ICR	IPA	SPU	MIS	NON	REC	PRE	OVG
inc-total	529	1189	160	63	24	0	23	942	282	718	36	16	79
perp-total	249	687	39	19	41	0	4	588	150	631	19	7	86
phys-tgt-total	255	280	26	12	28	1	10	214	189	1788	12	11	76
hum-tgt-total	594	236	82	42	28	1	38	84	442	2038	17	44	36
Matched/Missing	1627	614	307	136	121	2	75	50	1063	1203	23	61	8
Matched/Spurious	971	2392	307	136	121	2	75	1828	407	4601	39	16	76
Matched Only	971	614	307	136	121	2	75	50	407	629	39	61	8
All Templates	1627	2392	307	136	121	2	75	1828	1063	5175	23	16	76
Set Fills Only	778	333	177	35	74	0	7	47	492	538	25	58	14
String Fills Only	419	105	50	30	22	1	30	3	317	353	16	62	3

Table 3. TST3 (MUC4 Final Test) Summary Scores

Although our overall TST3 and TST4 scores clearly fell short of our goals, there are important comparisons to be made between TST2 and TST3. Most importantly, our recall scores made a definite improvement, as can be seen in the TST3 REC column (vs. the TST2 REC column) and on the Recall axis in Figure 1. This is due to improvements in the extraction of events and entities, from the text, at our knowledge representation level. (Some examples of this are given in the system summary paper in our discussion of Message 0048). Unfortunately, our precision did not significantly improve. This is in large part due to template overgeneration, which is caused by deficiencies in our event template merging. We are not yet properly merging event references across multiple sentences.

Although improvements in both recall and precision are required, we anticipate that first solving the overgeneration problem will give us a more accurate picture of how the system is really performing in terms of recall and

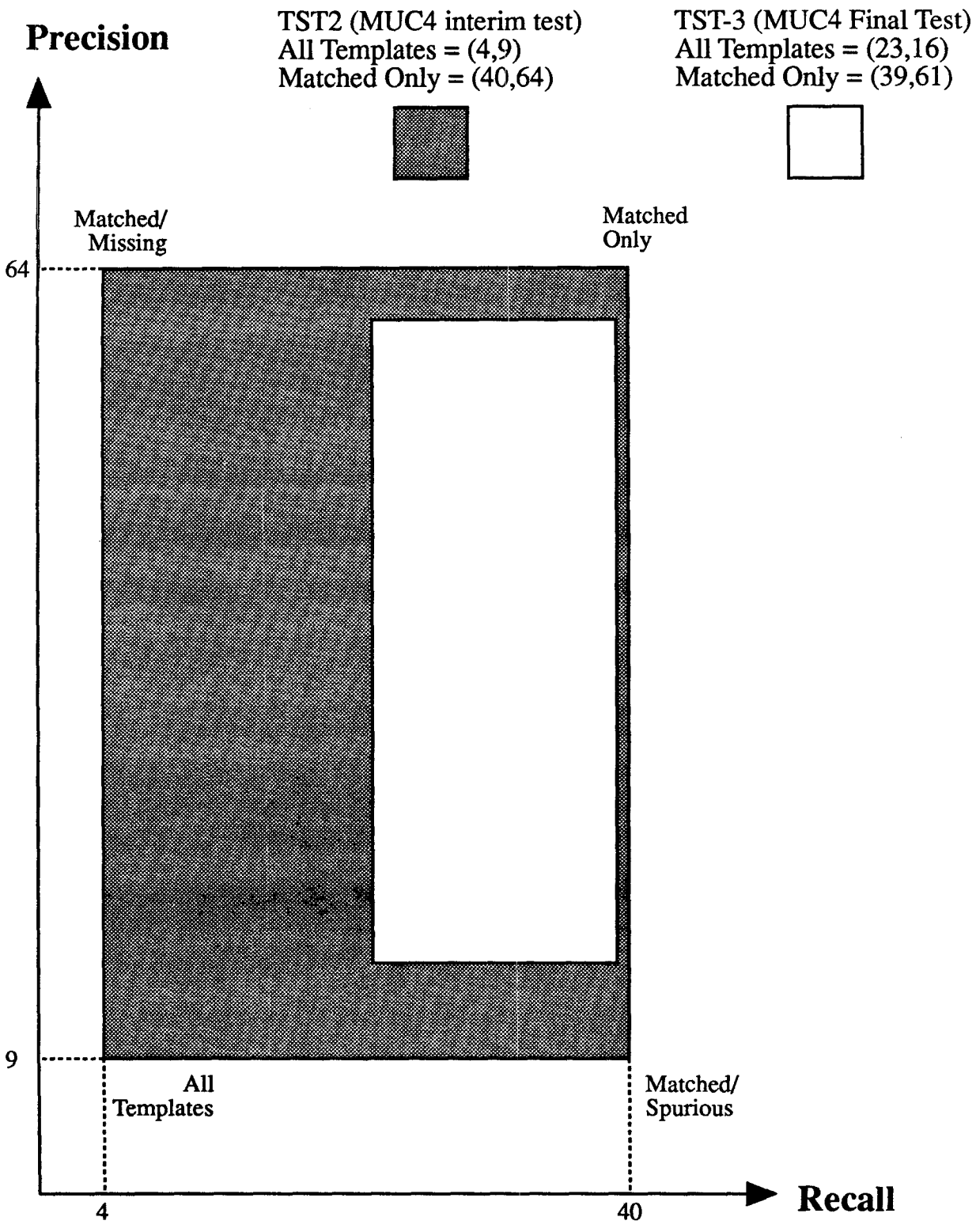

Figure 1: Recall and Precision (R, P) for TST3 vs. TST2

precision, and where additional work will produce the most significant improvement in system performance.

ALLOCATION OF EFFORT

Figure 1 of LSI's system summary in this proceedings presents an overview of the DBG system as configured for MUC-4. A new module has been added at the front end to select sentences of potential interest for the application. Work on our Principle-based parser has continued throughout the past year, extending the inventory of syntactic structures that can currently be handled.

The major MUC-4 effort was devoted to the lexicon (approximately 35%) and to the parser (about 20%), with other modules getting substantially less of the total effort, as shown in Table 4.

Sentence selector	10%
Lexicon	35%
Parser	20%
Functional Parse	5%
Discourse	5%
Frame System	5%
DBG Templates	10%
MUC Templates	10%

Table 4. Allocation of Effort for MUC4

LIMITING FACTORS

MUC is unfortunately a resource-limited undertaking for LSI; however, we did expend a significant effort on the lexicon and parser for MUC-4. Although LSI is a small company, we were able to devote these resources to MUC-4 in part due to the sponsorship of DARPA and BRL (see Footnote 1), and additionally, because the work was directly in line with our overall NLP objectives mentioned previously.

Limiting factors included all those on the list -- time, people, cpu cycles -- as well as the budgetary limits mentioned above. Knowledge was also a limiting factor in the sense that portions of the knowledge embedded in the system were not exploited, and other crucial knowledge was not added, due to resource limitations.

On the other hand, the amount of knowledge represented in the expanded lexicon is significant, so significant achievements are possible if limited resources are focused on particular problem areas.

TRAINING

During our preparation for MUC-4 testing, we were able to use the entire development corpus this year, and found it extremely valuable in our system development.

MODULE MOST OVERDUE FOR REWRITING

The code for our Lexical Unexpected Inputs/Word Acquisition Module (LUX/WAM), which deals with erroneous (e.g., misspelled) or new words is still the one which has gone for the longest period of time without rewriting or optimization of any kind. However, with our new, much larger lexicon, LUX/WAM was invoked far less frequently than during MUC-3 processing, and so was not really a significant factor in MUC-4.

A second module mentioned last year as a candidate for rewriting was LXI, the lexical lookup component. Some modification of LXI code was carried out to provide more efficient processing for MUC-4.

REUSABILITY

Throughout LSI's MUC participation, our goal has been to exploit this opportunity to achieve a generic, broad coverage, text extraction capability. To this end, with the exception of specific MUC-oriented parameters such as the names of critical events, the DBG system as configured for MUC is completely reusable in another application (and is in fact being used for all other NLP projects currently in house, including the NLP component of our voice translation testbed for English-->Spanish-->English). For example, the new sentence selection module

added this year can be used to search any text; only the tables containing MUC-oriented words that indicate critical event content are MUC-specific.

REFERENCES

[1] Berwick, R. C., Principle-Based Parsing, AI TR No. 972, June, 1987.

[2] Montgomery, C. A., Stalls, B. G., Belvin, R. S., and Stumberger, R. E. *MUC-3 Test Results and Analysis.* Proceedings of the Third Message Understanding Conference. Morgan Kaufmann Publishers, San Mateo, May, 1991.